

De la predicción a la acción

Por qué la IA probabilística necesita
límites deterministas

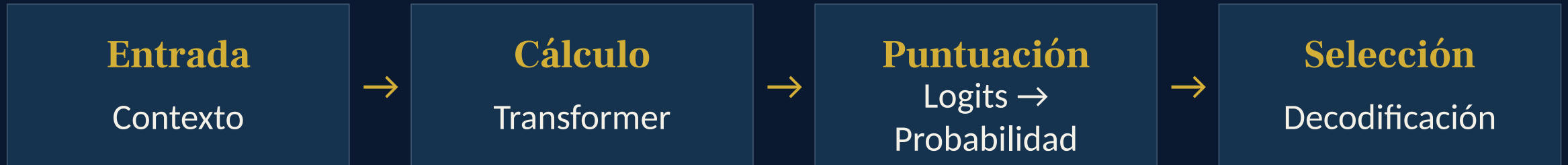
El modelo propone.

La arquitectura autoriza, valida y contiene.

Las personas conservan la autoridad.

¿Dónde aparece la incertidumbre?

Del contexto a la selección del siguiente token



El muestreo puede añadir aleatoriedad. Elegir siempre la continuación más probable tampoco garantiza un resultado correcto, seguro o alineado con la política.

La incertidumbre del modelo no es un fallo aislado: es una propiedad del sistema que debe ser gestionada con controles, validación y monitorización.

Del texto a las consecuencias

01 Generar

Una respuesta puede influir en una decisión.

02 Actuar

Una herramienta puede modificar datos, lanzar una transacción o interactuar con un sistema.

03 Delegar

Varios agentes pueden propagar un fallo, amplificar un error o encadenar acciones no previstas.

La frontera crítica aparece cuando una propuesta se convierte en una acción con consecuencias sobre datos, sistemas, procesos o personas.

Dos palancas de riesgo

Una heurística de ingeniería, no una fórmula completa

$$\text{Riesgo} \approx \text{Probabilidad} \times \text{Impacto}$$

Reducir probabilidad

Validación · pruebas adversarias · restricciones · hardening · monitorización

Reducir impacto

Permisos mínimos · topes · aprobación · reversión · contención

Heurística de ingeniería: además de probabilidad e impacto, el análisis del riesgo puede considerar exposición, detectabilidad, duración, reversibilidad y personas afectadas.

La envolvente determinista

El modelo

Razona, genera y sugiere.



El control

Verifica identidad, permisos, esquema, límites y políticas fuera del prompt.

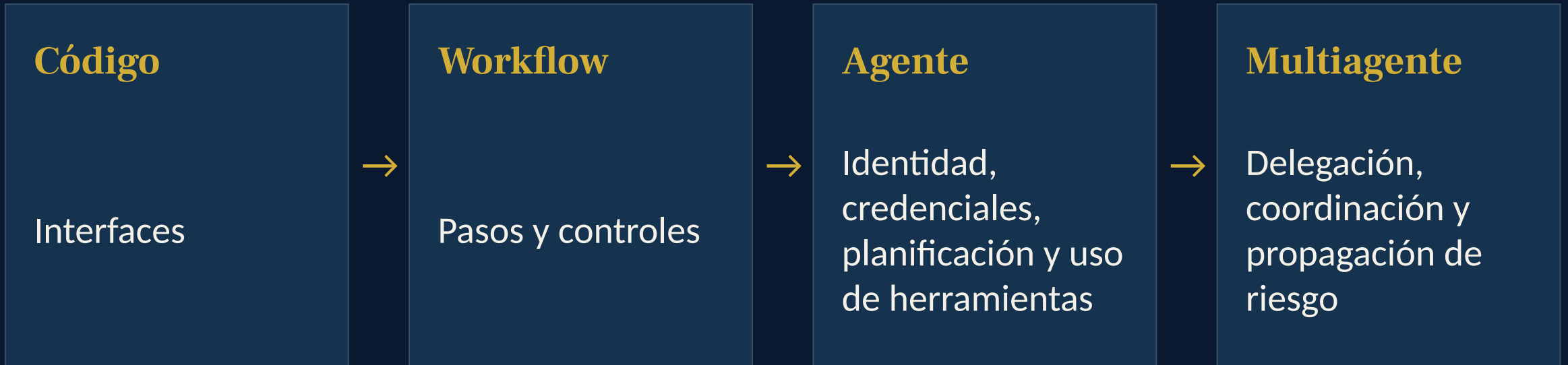


La ejecución

Acotada a herramientas autorizadas, trazables y dentro del radio de impacto.

El modelo sugiere; el sistema decide. Autoridad, datos, herramientas e impacto quedan acotados fuera del modelo.

Más autonomía, más contención



Delegar capacidad nunca debe aumentar la autoridad.

A mayor autonomía, mayor necesidad de mínimo privilegio, aislamiento, supervisión y límites verificables.

Separación de poderes para la IA

Legislativo

Define políticas, valores, límites y reglas de autorización.

Ejecutivo

El agente planifica y solicita llamadas a herramientas dentro del alcance permitido.

Judicial

Verifica, audita, registra evidencia y puede bloquear, escalar o detener.

Un agente no debería ser el único juez de sus propias acciones.

Analogía constitucional deliberada: en términos de ciberseguridad, equivale a segregación de funciones, control independiente y capacidad de intervención.

Gobernanza ♦ red teaming

El bucle de aprendizaje

Gobernanza de IA

Políticas y marcos
Matrices de riesgo
Reglas generales
Criterios de autorización

Red teaming

Ataques adversarios · Pruebas de estrés
Prompt injection · Tool abuse
Policy bypass · Exfiltración de datos

CICLO

Hipótesis → Despliegue → Falsación → Hallazgo → Regla determinista → Test

Cada hallazgo adversario pasa a ser una regla determinista y un test de regresión.

La misma IA, dos resultados

Demostración en dos actos · misma tarea · [Ver vídeo](#)

Sin control

El agente intenta ejecutar la acción propuesta con capacidad potencial de error, abuso o escalado no previsto.

Con runtime assurance

Intercepta → valida → decide
Bloquea, autoriza o escala
Deja evidencia y registro

**El mismo modelo. La misma tarea.
Distinto poder de decisión.**

La seguridad no depende solo de la inteligencia del modelo, sino del plano de control y de las políticas que gobiernan su ejecución.

Gobernanza como control en ejecución

Modelo probabilístico



Política, identidad y autorización



Ejecución acotada



Evidencia, monitorización y prueba adversaria



Respuesta a incidentes · revisión de límites · mejora continua